DOCUMENT RESUME

ED 450 818 JC 010 192

AUTHOR Luan, Jing; Willett, Terrence
TITLE Data Mining and Knowledge Management: A System Analysis for Establishing a Tiered Knowledge Management Model.
INSTITUTION Cabrillo Coll., Aptos, CA. Office of Institutional Research.
PUB DATE 2000-00-00
NOTE 15p.
PUB TYPE Reports - Descriptive (141)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Community Colleges; *Data Analysis; *Educational Research; Educational Researchers; *Research Tools; Two Year Colleges
IDENTIFIERS Cabrillo College CA; Knowledge Management

ABSTRACT

This paper discusses data mining--an end-to-end (ETE) data analysis tool that is used by researchers in higher education. It also relates data mining and other software programs to a brand new concept called "Knowledge Management." The paper culminates in the Tier Knowledge Management Model (TKMM), which seeks to provide a stable structure with which to organize the plethora of established and nascent technologies. Data mining is a knowledge discovery process to reveal patterns and relationships in data via high-powered data modeling procedures. The field is in the process of being harmonized with statistics to provide researchers with a richer and more unified palate of analysis tools. The birth of data mining, however, has not completed the road map for research in higher education. With the development in data warehousing and data mining, the landscape for knowledge management has greatly changed. After extensive research and based on actual experience, a model for managing knowledge for research and planning is proposed to be the Tiered Knowledge Management Model (TKMM). A roadmap like TKMM may help guide the efforts for researchers to update their skills and choose the right tool. For example, the Project Management model explains what tool is best for which project. Addendum describes five steps to successful data mining. (Contains 13 references.) (JA)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 450 818

Data Mining and Knowledge Management

A System Analysis for Establishing a Tiered Knowledge Management Model

Jing Luan, Ph.D.
Terrence Willett, M.S.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

John Nurd.

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Cabrillo College
Office of Institutional Research
6500 Soquel Drive
Aptos, CA 95003

TC010192

BEST COPY AVAILABLE

Data Mining And Knowledge Management

A System Analysis for establishing a Tiered Knowledge Management Model

Jing Luan, Ph.D.
Terrence Willett, M.S.

Researchers in general are living in an era of great advancement in technology. This brings significant pressure to update skills and analysis tools. Researchers in higher education in particular have gradually begun to realize that yesterday's modus operandi is dated, as technology is knocking on their office doors with promises of efficiency, speed, and convenience. In response, many researchers have opened their doors to new software and new technology. A recent Intellisolve Group's literature reflected on the patterns of adopting technology-using terms such as early adopters, early majority, and laggards (Intellisolve Group, 2000). The early adopters have left a trail of confusion to the early majority and laggards who follow. The road is littered with names such as XML, ASP, and JSP, to name a few with new acronyms being born each day. Some of them are being recycled, such as ASP, which has had at least five uses. Then there are data mining, data-marts, OLAP, and other concepts and tools being introduced with each new innovation.

While it is not the purpose of this paper to address them all, it is the intent of the authors to focus on data mining--an end-to-end (ETE) data analysis tool. Then the authors will proceed to relate data mining and other software programs to a brand new concept called "Knowledge Management". The paper culminates in the Tier Knowledge Management Model (TKMM) that seeks to provide a stable structure with which to organize the plethora of established and nascent technologies.

Introducing Data Mining

Data mining cannot be discussed without first relating it to datawarehousing. A datawarehouse is a "subject oriented, integrated, nonvolatile, time variant collection of data in support of management decisions" (Corey & Abbey, 1999). Many researchers have established datawarehouses on their campuses. If they haven't yet, they might have datamarts.

The difference between a datawarehouse and a datamart lies in the size and purpose. Datamart means a specially constructed database for a limited research study, such as the study of transfer students' course taking pattern, whereas datawarehouse comes from an OLTP system such as Peoplesoft or Datatel. The datawarehouse is comprehensive in nature, which may become too large for a study that only uses a few fields of data without the introduction of another layer of data management (described later as the Middle Tier). Datawarehousing technology has drastically expanded the amount and type of data available for analysis, therefore making data mining a possibility.

Data mining is a knowledge discovery process to discover patterns and relationships in data via high-powered data modeling procedures (De Veaux, 2000). Data mining is about discovering rules, associations, and likelihoods of events by looking backward at

data in its raw and longitudinal form. The power delivered by datawarehousing to data mining software has broken the mold of the traditional statistical methodologies revered by generations of analysts (Mena 1998). Trained in the theories of sampling and the sutra of hypothesizing, statisticians have been practicing their trade with pre-conceived notions, or a-priori hypotheses. If the traditional methods can be viewed as top down, data mining is truly bottom up. Research questions do not begin with "what *if*", instead, they begin with "what *is*".

Data mining has been recently discovered by academia but was first put to full use by the Fortune 500 who have since benefited tremendously. Data mining was behind numerous successful market campaigns and quality assurance (QA). Below is a comparison chart depicting some of the core questions most often used in the business world and their analogs in higher education:

Bottom-line Questions in the Business World	**Counter-part Questions in Higher Education**
Who are my most profitable customers?	Who are the students taking most credit hours?
Who are my repeat website visitors?	Who are the ones likely to return for more classes?
What clients are likely to defect to my rivals?	What type of courses can we offer to attract more students?

Exploring Data Mining for Higher Education

In a traditional approach to science, the researcher would make specific a-priori hypotheses and then test them using data collected specifically to test these hypotheses. For instance, a researcher might propose that the probability of a transfer student graduating can be predicted from GPA at time of transfer, GPA at the transfer institution, ethnicity, and field of study. Logistic regression or some other analysis would test this model for significance. In a strict analysis, we would be left with our hypothesis either being supported or not. Further testing after the fact would encourage the discovery of spurious findings unless care was taken to protect against alpha inflation and even then one could find grounds to object to such post-hoc "fishing".

In an exploratory setting where we are searching for models rather than testing hypotheses, the researcher would retain variables that predicted a significant proportion of the variance and discard those that did not. The researcher could choose to alter the model by adding previously ignored variables such as mean unit load per term or make finer adjustments. We might also wish to test for interactions or control for confounding variables. Perhaps a variable excluded at one step becomes important in conjunction with a variable added later in our discovery process. A researcher in such a procedure would have to be highly adept at statistics and data manipulation and spend a great deal of time and effort in this search for models and patterns and still might miss a parsimonious model.

The researcher in this example is data mining but is doing so "by hand". It is the introduction of the computer that makes even this manual level of data mining possible. One might even joke that it was the thought of performing repeated analyses literally by hand that motivated the tradition of the a-priori hypothesis during the pre-computer development of statistics. The advent of the computer and its exponential increase of processing and storage capabilities and reduction in price revolutionized numeric analysis and to date culminates in programs such as data mining software. Rather than laboriously test a few variables at time and alter iteration parameters, the researcher assigns variables independent ("in") or dependent ("out") status, chooses a mining method, and lets the program come up with a set of rules that governs the situation. The set of rules is clear even to those not familiar with statistics. For instance (using imaginary data), a simple model might say:

If transfer GPA > 3.0, then graduation = yes (confidence = 0.87)
If transfer GPA ≤ 3.0 and
  If university GPA > 3.5, then graduation = yes (confidence = 0.97)
  If university GPA ≤ 3.5, then graduation = no (confidence = 0.77)

This result is easy to obtain and has a clear interpretation. The data mining program has processed a large number of variables and considered them all in the analysis and has tested its own predictions to improve its model and provide empirically based confidence levels.

An analyst might look with some suspicion on such numeric magic (Elder and Pregibon 1996). Many would be wary to trust an analysis that we couldn't replicate by hand or at least by a proven statistical program. However, the mining program is based upon machine learning algorithms similar to our own mind's discovery process of trial and error and derives it mystique from the ability of the computer to calculate more quickly than most humans.

Hosking et al. (1997) point out that in statistics, we often work with fixed conceptual and hypotheses spaces where we select a set of parameters and a model to test our hypotheses. In data mining, we also have a fixed conceptual space but the hypothesis space is left to the learning algorithm, which attempts to create a model with a minimum of prediction error. Data mining also relies less on assumptions about data distributions and generally results in more complex models judged not on how well they support theory but on how well the model generalizes to new data (Bengio et al. 2000). In general, statistics has been developed for testing simple models where variables of importance are selected by educated judgment or by theory (Table 1). Data mining has been designed to operate on large data sets containing numerous variables with unknown or complex relations. These approaches can also be combined. For example, a linear combination derived from a discriminant analysis can be included as a variable in a data mining effort. Exploratory data analysis (EDA) techniques such as factor analysis blur the distinction between statistics and data mining and help illustrate how these approaches are not by any means mutually exclusive but can help confirm results in some cases. A major difference between EDA and data mining is that EDA, coming from a statistical tradition, often relies on specific distributions and covariances of variables

while data mining techniques, originating in a machine learning and artificial intelligence context, use algorithms designed to seek patterns.

Table 1. Non-exhaustive general comparison of statistics with data mining from the perspective of a social scientist.

Inferential Statistics	Exploratory Statistics	Data Mining
t-test, ANOVA, regression, Chi-square	Factor analysis	neural networks, Kohonen networks, C5.0 rule induction
a priori hypotheses	factors	generated rules
simple models	simple to complex models	complex models
small data sets	large data sets	very large data sets
variables of known importance	variables with unknown relevance	variables with unknown relevance
"familiar" formulae	less "familiar" formulae	"Black Box"
distributional assumptions	distributional assumptions	learning algorithms
significance	significance	accuracy and generalizability

De Veaux (2000) listed five data mining models: Descriptions, Classifications, Regressions, Clustering, and Association. He argued that OLAP (On-line Analytical Processing) and query software, such as BrioQuery, fall in the category of Descriptive. Classifications work with sets of categorical or discrete values while Regression analyzes continuous values. Clustering discovers groupings within data sets and Association elucidates hidden relations among data elements. Some data mining techniques touch on more than one model. Examples of techniques include neural nets such as Kohonen, Kmeans, and Nearest Neighbor (KNN) that are often used for clustering. For classification and association through decision trees, some of the most often used are: CART (Classification and Regression Trees), CHAID (Chi-square-Automatic-Interaction-Detection), and GRI (Generalized Rule Induction), and C5.0.

Data mining is a growing field expanding into many disciplines with large and complex data sets including business, education, cladistics, atmospheric science, astrophysics, and genetic research. New techniques are being developed continuously in response to demands from particular research areas and opportunities offered by improved processing technology. Data mining is also in the process of being harmonized with statistics to provide researchers with a richer and more unified palate of analysis tools.

Exploring Data Mining in a Higher Education Setting (A Case Study)

Transfer is one of the three missions for community colleges. For many years, transfer reporting has been predominantly a numeration of the total numbers of students moving from community colleges to universities, as has been the case in California. No meaningful research can be conducted on these transfer students because of lack of unitary data. In 1998, three separate joint partnership agreements resulted in course level transfer student data being provided to Cabrillo College's Office of Planning and Research (PRO). University of California Santa Cruz (UCSC), San Jose State University (SJSU), and California State University Monterey Bay (CSUMB) have provided longitudinal data on the students who have transferred from Cabrillo to their institutions in the past five years. This partnership represents ⅔ to ¾ of our transfers in any given year. The files contained information on their course enrollment, majors, and their graduation information for the purpose of identifying the education outcomes former Cabrillo College students obtained after they left Cabrillo. This type of longitudinal course level research has gone beyond simple head counting. It answers many questions that faculty at community colleges have been asking, such as what are the potential factors influencing students' educational outcomes after they leave.

Define the Research Question(s) Well

A policy question on transfer can be "what measures can a community college take to increase transfer rates?" This policy question is translated into the following research question: what factors to what degree are influencing the student in his/her transfer behavior and based upon this what is the profile of a transfer student regarding the student's course taking pattern, course outcomes, demographics, and his/her educational outcomes in the transfer institution that can, in turn, be used to modify the community college curriculum?

A test set was created using data from all Cabrillo College students enrolled in the 1997-1998 academic year (please refer to the addendum for the five essential steps in building data sets for data mining). The study chose Clementine as the data mining software (please also refer to the addendum for data mining tools). Further, the main data mining technique was C5.0, a decision tree model. The data file is a single occurrence file (one record per student) restricted by term, which was a product of a one to many and many to many relational database querying. No restrictions existed on how many variables there were to enter into the data mining data file. As an example, some of the variables are listed as follows:

- Total Units Attempted, Units Earned, and Grade Points
- Grade Point Average
- Total Transfer, Vocational, and Basic Skills Courses Taken
- Educational Goal
- Demographics: Age, Gender, Ethnicity, Disability, Education, City, Major
- Financial aid and AFDC status
- Transfer Status in Fall 1998

This is not a complete list of the variables, nor is it considered a large pool of variables, as most data mining software is capable of handling very large numbers of variables at one time.

Interpreting Data Mining Outcomes

The research explored what factors are associated with students who either transferred or did not in Fall 1998. Decision Tree C5.0 in Clementine uncovered a long set of rules. Some are highlighted below with confidence measures in parentheses.

For those with less than 71 units attempted, will transfer if

Educational goal is a vocational AA/AS (56%)

If educational goal is AA/AS and receive financial aid (67%)

If educational goal undecided and more than 23 transfer courses (79%)

If educational goal is transfer with AA/AS:

Unknown majors over 20 years old (87%)

If educational goal is to form career interests and more than 2 basic skills courses and receive financial aid (100%)

If GPA is greater than 3.2 (90%)

For those with more than 71 units attempted will transfer if:

Not Native American (100%)

White and either

High school diploma or not and more than 9 transfer courses (100%)

AA/AS or higher and no vocational courses (100%)

The findings point out that a student's stated educational goal was important in predicting transfers, as was the presence of receiving financial aid in some cases. Ethnicity appeared important only for students with a large number of units attempted. This profile indicates to campus marketers, policy makers, and creators of curricula who is currently being served, who needs outreach services, and what factors appear critical for successful transfer outcomes. A college could use such findings to guide counseling and other matriculation efforts to certain groups and to justify financial assistance to students.

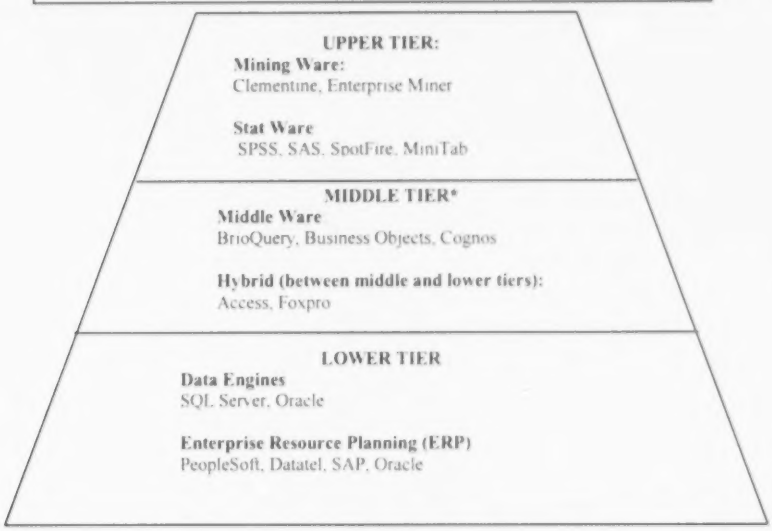
We can also compare results from data mining to statistical analyses. For example, a Neural Network algorithm identified the following five variables as being most important in predicting transfer status: Major, educational goal, number of transfer courses taken, number of units attempted, and age with major being by far the most important. A stepwise logistic regression entered the variables number of transfer courses, educational goal, units attempted, vocational courses, residency, GPA, age, major, and education in that order. There is much overlap in variables but differences occur in the influence attributed to each possibly due to the interaction of the nature of the variable and the model characteristics.

Note that some variables are noteworthy by their exclusion. Neither gender nor disability appeared in any of these models nor were there significant differences between groups in proportions of transfer status using Chi-square analysis (with the exception that females with transferring as an educational goal were significantly more likely to transfer than males but with a small effect size). This suggests that matriculation efforts to give equal opportunity to transfer to gender and disability groups do not appear to need major adjustment based upon our data set. At this point, it is prudent to remember that these rules and models do not necessarily generalize to other academic years or institutions.

From Data Mining to Knowledge Management

The birth of data mining appears to have completed the road map for research in higher education. From this moment on, researchers have an entire set of tools for their trade on their desktop. However, in a fast changing and technology driven world, one can easily be overwhelmed by a number of factors. For example, in the late '80s, terms like Decision Support Systems (DSS) and Executive Information Systems (EIS) became fashionable to describe the process of using data to guide our decisions (Intellisolve Group, 2000). However, higher education is not about making decisions only and Decision Support is too closely associated with administration and management, unfriendly terms to some faculty and students. Since higher education is about the creation, transformation and transmission of knowledge (Laudon, 1999), an appropriate and commonly accepted term is now "Knowledge Management", the process for which is called "Knowledge Discovery". With the development in datawarehousing and data mining, the landscape for knowledge management has greatly changed. It is now time to take a bird's-eye view of the playing field. After extensive research and based on actual experience, a model for managing knowledge for research and planning is proposed to be the Tiered Knowledge Management Model (TKMM) (Figure One).

Figure One. Topography for a Tiered Knowledge Management Model



* Web based client side (end user, thin client) OLAP software now includes JSP, Dreamweaver, Coldfusion, etc.
Note: this typography for TKMM is for illustrative purpose only. Not all products by all vendors are displayed nor are any products shown necessarily endorsed.

At the bottom of the Lower Tier, Enterprise Resource Planning (ERP) systems, such as Peoplesoft or SAP, are the source for most of the data that researchers need, short of surveys and other supplemental sources. These systems are on-line transaction processing (OLTP) applications that maintain the most scattered and fragmented relational data files. The data from OLTP should be first staged and denormalized before moving into a datawarehouse where SQL servers and their ilk unleash their power. The Middle Tier houses programs that act between the servers and the clients for data querying and reporting. Some call them Middleware, as in a distributed network-computing environment. The Upper Tier of this model points to traditional statistical packages as well as the most exciting tool: data mining software programs.

What's the Significance of TKMM?

Coming together as a knowledge management model, TKMM holds significant implications to researchers in the areas of securing funding, updating knowledge, managing the office, and understanding the relationships between research and other technology intensive departments on campuses. A road map like TKMM may help guide the efforts for researchers to update their skills and choose the right tool. Specifically, the implications are in the following areas:

- Project Management-This model explains what tool is best for which project, e.g., for real-time query as in census FTES reporting, you need OLAP tools; for building data sets for Statware or Miningware, you need relational database querying tools.

- Skills Update-This model describes the relationship of the software programs in each tier and the level of knowledge needed for each, e.g., to be comfortable with the Middle Tier, you should know Javascripting for Brio, web based querying, and SQL (Structured Query Language)
- Managing the Office-The model helps you identify on which tier you have the strength and for which tier you need to work with other departments. It helps you determine your SOP (Standard Operating Procedure), e.g., understanding how data is processed into datawarehouses and by whom, and what you should expect from your data processing department.
- Resource Planning-This model guides the planning and allocation of resources for data and for research, i.e., using the Total Cost Ownership model (TCO), you can successfully argue what software and hardware to invest in and when to upgrade them.

For research and planning to be useful and for all of us to be successful, the current knowledge base must be updated with new technology and TKMM is a good starting point.

Conclusion

Insights from data sets and variable lists previously seen as unwieldy and chaotic can be drawn out with data mining and developed into the foundations for program planning or to help focus research efforts. For e-commerce, data mining helps with individualized marketing; for the insurance industry, fraud detection; for manufacturing, most efficient approach; and for education, enrollment management. The use of data mining is not just limited to these few examples, but what can be uncovered via a comprehensive and automated data mining process can mean millions of dollars for any given user. A one-percentage point (1 %) may mean $500,000 for a typical college of 20,000 students. Data mining conducted for alumni donations may correctly pinpoint the right donors and the right target amount. This both saves campaign costs and increases the campaign achievement. Data mining conducted to predict the likelihood of an applicant's enrollment following their initial application may allow the college to send the right kind of materials to the potential student and prepare the right counseling for him or her. The potential of data mining in education cannot be underestimated.

To successfully engage in knowledge discovery, the TKMM model can help researchers determine what tools best fit for a particular project (Project Management), what skills are a must for surviving in a fast changing world of technology (Skills Update), which tier in the model contains most of the human capital of the research office and on which tier the office needs cooperate with other departments (Managing the Office), and what to buy and when to buy (Resource Planning).

Addendum

Data Preparation and Validation

Data mining software is not built for, or has not evolved into, working directly with relational databases or conducting SQL queries. Referring to the TKMM (Figure One), data mining relies on the work from lower tiers in TKMM. The authors of this paper propose the following five steps to successful data mining. The first four are all related to the steps before data is brought into a data mining environment.

Step One – Move data into a datamart. A researcher may need to first build a datamart to house the unitary relational data files. A datamart will help avoid interfacing with a datawarehouse. Depending on the size of the datawarehouse and the technical skills of the researcher, this may or may not be totally necessary.

Step Two – Querying data for building a flat file. This step calls for skills in SQL commands and familiarity with the four types of joins: basic join, left-outer join, right-outer join, and full-join. There needs to be one occurrence per record with multiple occurrences converted into fields (i.e., recode all courses taken by a students into types with each type occupying a field that holds the number of courses taken within type), or multiple values aggregated (i.e., units counts for courses collapsed in one value). It is strongly recommended that the researcher save the SQL in case there is a need to redo the flat file or slightly modify the SQL for other uses. BrioQuery (*.bqy files), Foxpro (*.qpr files) and even SPSS (*.sps scripts) can be useful, although SPSS works with data transforming only within a flat file.

Step Three – Data visualization. This means both examining frequency counts as well as generating scatter plots, histograms, and other graphics. A graph is the best spokesperson for a correlation estimate. This step gives the researcher the first impression of what each of the data fields contains and how they may play out in the analysis. More often than not, the researcher needs to be very familiar with his or her data elements and this step is a sure way to guarantee that.

Step Four – Data validation. This step may take place simultaneously with Step Three. Depending on the type of query software, some data elements may have been extracted erroneously. For example, Foxpro does not like multiple joins for two files. The second join condition may not run or may bring back a completely wrong set of data all displayed in the correct field with the correct values.

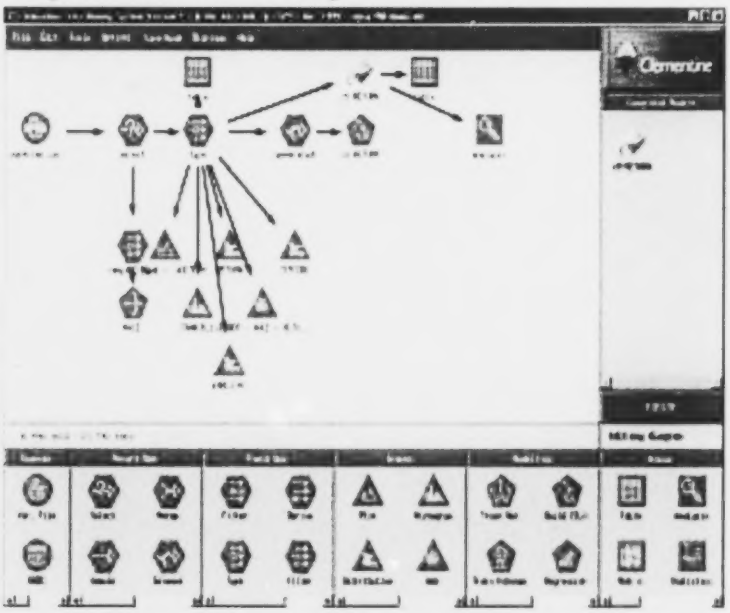
Step Five – Mine your data!

In a simple data mining task for the case study in this primer, a number of issues confronted the researcher even before the data could be brought into Clementine. These issues helped to demonstrate the importance of data preparation. The issues manifested themselves in several areas. They were (1) compiling a data set from one to many and many to many relational databases, (2) converting variables that were not meaningful or far too inclusive, i.e., 27 different ethnicity categories, (3) delineating courses into

transfer, vocational, and basic skill course count from course names, (4) resolving missing values (data imputation), (5) deleting data entering errors or system determined values. Frequency counts also uncovered other minor problems that would have remained unnoticed until outrageously alarming findings came out of the analysis. It took several days to generate the final file while it took only an hour to complete the first run using C5.0. There is truth in the statement that researchers should spend 95% of their time preparing the data. Other lessons learned are the need for carefully defining the research question and the benefit of knowing the data element dictionary (DED) well.

Choose the Right Tool

Figure Two. Clementine Working Model:



Some in the industry call data mining software packages "tool boxes", and the modeling techniques "tools". Others view data mining software as an end-to-end solution. Clementine (Figure Two), for example, is a premier data mining software added to the ammo of SPSS in the mid '90s. SAS has just recently rolled out Enterprise Miner and SGI MineSet. Clementine has a number of features that users will come to appreciate. It is laid out differently then the typical Windows product, but that may be where some of its strength lies. The program is very Graphical User Interface (GUI) intensive. The lower part of the window contains various toolboxes with graphing and analysis functions. The larger space, called a pane, is the workbench. The data analysis is represented by a visual basic programming flow chart called a stream. The data typically flows from a data source node to a data manipulation node to a modeling node. The most often used modeling techniques for educational research are neural nets and decision trees. Neural nets deal with numerical data exclusively while decision trees can incorporate categorical data. Some of the most often used neural net techniques are Kohonen, Kmeans, and Nearest Neighbor (KNN). Some of the most often-used decision trees are: C5.0, CART (Classification and Regression Trees), CHAID (Chi-square-Automatic-Interaction-Detection), and GRI (Generalized Rule Induction).

BEST COPY AVAILABLE

References

_____ (1998) Clementine Data Mining Systems, Integral Solutions Limited (ISL), 1998

Bengio, Y., Buhmann, J.M., Embrechts, M., Zurada, J.M. (2000) Introduction to the special issue on neural networks for data mining and knowledge discovery. Transactions on Neural Networks. 11.3: 545-549.

Corey, M., Abbey, M. et al. SQL Server 7 Data Warehousing. Microsoft Corp, 1999

De Veaux, R. (2000). Data Mining What's New, What's Not. Presentation at the Data Mining Workshop, Long Beach, California.

Elder IV, J.F. & Pregibon, D. (1996) A statistical perspective on knowledge discovery in databases. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy, eds. Advances in knowledge discovery and data mining. MIT Press: Menlo Park, CA. 83-116.

Hand, D. (1999) Statistics and Data Mining: Intersecting Disciplines. SIGKDD Explorations. ACM SGKDD. Volume 1, Issue 1 (page 16).

Hosking, J.R.M., Pednault, E.P.D., & Sudan, M. (2000) A statistical perspective on data mining. Future Generation Computer Systems. 13: 117-134.

Laudon, K., Laudon, J. (1999) Management Information Systems-organization and technology in the networked enterprise. Prentice Hall, New Jersey.

Luan, J., & Torpy, Ed., (2000) Proceedings/Presentation at the 40th AIR Forum, Cincinnati, OH.

Luan, J., Holbert A., & Satren M. et al. (1996) Using A Datawarehouse to Conduct Longitudinal Tracking of Watsonville Students. Proceedings at the 35 Annual RP Conference, Berkeley, CA.

Mena, J. (1998) Data-mining FAQs. DM Review, January 1998.

Perry, P. (1999) Microsoft SQL Server 7.0 and Brio Enterprise Fuel California Community Colleges' "Student Right-to-Know" Program. Industry Solutions Vol. 4. Microsoft, Redmond, WA

Pickering, C. (2000) They're Watching You. Business 2.0 (page 135 – 136) Feb, 2000.

Useful Websites:

For statistical terms:

http://www.statsoft.com/textbook/stathome.html

For machine learning and data mining terms:

http://www.cs.sfu.ca/people/GradStudents/melli/MD_Terms.html

For searching papers and new ideas:

http://www.spss.com/searchaction.cfm

Cabrillo College Planning and Research Office

http://www.cabrillo.cc.ca.us/oir

About the Authors:

Jing Luan is Director of the Planning and Research Office at Cabrillo College. A veteran in his field, he is a published scholar and a practitioner in a variety of higher education areas. He currently chairs the California Community College Decision Support System Datawarehousing Project and is the President-elect of his trade organization, the RP Group. He can be reached at jing@cabrillo.cc.ca.us

Terrence Willett is a Researcher and Technician at the Planning and Research Office at Cabrillo College. An environmental scientist, he is skilled in using multivariate analyses to model complex and difficult to measure interactive systems. He can be reached at tewillett@cabrillo.cc.ca.us